Alternative Approaches To Knowledge Discovery in Official Statistics

Thanyani Alpheus Maremba*
Statistics South Africa, Pretoria, South Africa ï thanyanimar@statssa.gov.za

Abstract

Statistical agencies accumulate a large amount of data from recurring surveys and administrative records over a number of years. The data is analysed periodically and results are provided to the general public. With the amount of data produced on a regular basis, historical data become more difficult to analyse using standard systems designed for the statistical publications. Users both internal and external to the statistical agencies usually have requests for historical data that is not published. Making information more broadly and easily available to more users throughout the agency and beyond the agency to stakeholders is becoming of a more strategic importance. There are techniques that are available and are considered by Statistics South Africa in an effort to enhance analytical capability for both technical and non-technical users. The techniques considered include online analytical processing (OLAP), dimensionally aware relational schemas, geographic data mining and knowledge discovery. These approaches will help the agency to make the most out of the available data. In order to achieve the best results, development of the statistical indicators for decision support aimed at policy decisions or quality management decisions as well as creation of data marts become essential.

Keywords: Statistical indicators; knowledge discovery; DSS; OLAP

1 Introduction

The alternative approaches to knowledge discovery draws from the areas of statistics, machine learning, data mining, decision support systems, operations research as well as data visualization. There are lessons learnt from Business Intelligence (BI) to use internal data in building intelligence. BI tools are commonly used in profit driven organisations such as banks, insurance companies, etc. These techniques are used to give companies competitive edge by using the data from within to improve profits. Statistics South Africa is quality driven organisation. BI techniques can be used to enhance quality of our products by providing relevant quality indicators. Decision makers will put measures that will enhance quality of statistical data and related products.

We apply this principle in one of Stats SA survey that is Quarterly Labour Force Survey (QLFS). We identify data needs from the regular requests by the management such as slippage rates, response rates, imputation rates, etc. We create data marts from the series of surveys or by quarters in the case of QLFS with specific topics such as slippage rates. The data marts are created with imbedded multi-dimensional structures, they are updated by appending every new quarter' data. To view these multi-dimensional data sets we implement the view using Online Analytical Processing (OLAP) system.

Emphasis in Stats SA survey is on the publication targeted to the public. Survey quality indicators such as response rates, slippage rates as well as measures of precision are produced with particular publication. The challenge is to study how surveys are performing overtime. We currently do not have quality indicators available to share with the relevant users in a single view from previous surveys.

To meet the request for the required indicators overtime we generate adhoc reports using SAS to extract data from various sources on the SAS server and tabulate it according to the request. The process of generating those reports is time consuming. Survey methodologists can save time spent in generating reports by implementing the proposed OLAP system.

In light of the current applications of visualisation that are applied in the Statistics South Africa survey data we explore the alternative approaches to knowledge discovery. The paper will look at the Decision support systems (DSS), knowledge discovery, geographic knowledge discovery (GKD) and spatial data mining, Visualisation techniques, spatial representation of conceptual multidimensional models. In addition we demonstrate the creation of data marts and the use of OLAP system as a method of information delivery.

2 Decision support systems (DSS)

The paper emphasises importance of a structured decision support system and the knowledge discovery techniques that can be used to provide information, knowledge and intelligence to the organisation. According to Khorshid (2004) a computer-aided decision support system (DSS) is conceptually composed of four components; (i) Database management capabilities with access to internal and external data, information and knowledge; (ii) Modeling functions accessed by a model management system; (iii) A powerful yet simple user interface design that enables interactive queries, reporting, and graphing functions; and (iv) A decision-makerôs own insights. By integrating various modeling capabilities, a DSS can be successfully used to support problem solving, policy testing, scenario simulation and strategic planning. Given the central role of mathematical models in explaining and interpreting the behavior of socioeconomic systems, DSS would represent a useful tool of analysis in this respect. In 1985, the cabinet of Egypt established the Information and Decision Support Centre (IDSC) whose mission is to provide information and decision support services to the cabinet for socio-economic development Kamel (1997). We will further enhance our decision support systems by incorporating knowledge discovery tools into our decision making process.

3 Knowledge discovery

Miller (2004) has outlined the process on knowledge discovery from databases as; background knowledge, pre-processing, data mining and knowledge construction. Alfred (2005) described data mining as the core task in the process known as Knowledge Discovery in databases. It consists of applying computational techniques to extract useful pattern or knowledge from the given data, typically expressed in the form of a predictive or descriptive model. Knowledge discovery in databases (KDD) was initially defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data by Frawley et. al., (1991). In 1996, KDDôs definition has been further revised as follows by Fayyad et. al., (1996), KDD is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern or knowledge in data. It is estimated that about 80% of data stored in databases has a spatial or location component, the location dimensions have been widely integrated in DWs and in OLAP systems Malinowski (2004). As a result we further look at concept of geographic knowledge discovery and spatial data mining to study the patterns that are related to location.

4 Geographic knowledge discovery (GKD) and spatial data mining

Geographic Knowledge Discovery (GKD) is based on a belief that there is novel and useful geographic knowledge hidden in the unprecedented amount and scope of digital geo-referenced data being collected, archived and shared by researchers, public agencies and the private sector. Geographic knowledge discovery (GKD) is the process of extracting information and knowledge from massive geo-referenced databases. The nature of geographic entities, relationships and data means that standard KDD techniques are not sufficient Shekhar et al. (2003). Specific reasons include the nature of geographic space, the complexity of spatial objects and relationships as well as their transformations over time, the heterogeneous and sometimes ill-structured nature of geo-referenced data, and the nature of geographic knowledge Miller (2004).

According to Miller (2004), spatial data mining process is made of spatial classification, spatial association, spatial classification and prediction, spatial clustering, spatial outliersô analysis. A spatial database is organized in a set of thematic layers. A thematic layer is a collection of geographical objects that share the same structure and properties. A theme can represent a road network, and another can represent towns. These allow to selectively using the relevant themes for a specific purpose.

Spatial relationships represent an essential characteristic in real world. They show spatial influences between entities. Indeed, observations located near to one another in space tend to share similar attribute values. This is known as one of the ñ1st law in geographyò specified by Tobler. Moran has defined a measurement of spatial auto-correlation between nearby data since 1948. Anselin has refined this in local auto-correlation indices qualifying the correlation of each entity value with the values of its neighborhood. Another technique investigated is spatial representation of conceptual multidimensional models which can independently be used to reveal knowledge from data and it can also be used in collaboration with data warehousing through OLAP systems as well as in the process of data mining.

5 Spatial representation of conceptual multidimensional models

Malinowski (2004) defined a Data Warehouse (DW) as a collection of subject-oriented, integrated, non-volatile, and time-variant data supporting management's decisions. Online Analytical Processing (OLAP) systems allow decision-making users to dynamically manipulate the data contained in a DW. OLAP systems use a structure called a cube, based on dimensions, measures, and hierarchies. Hierarchies allow the user to see detailed as well as generalized data using the roll-up and drill-down operations. Further, the slice and dice operations allow to select a portion of the data based on specified values in one or several dimensions. Since it is estimated that about 80% of data stored in databases has a spatial or location component, the location dimensions have been widely integrated in DWs and in OLAP systems. However, these dimensions are usually represented in an alphanumeric, non-cartographic manner, since these systems are neither able to store nor to manipulate spatial data. The management of this kind of data is usually carried out by Spatial Databases (SDBs) or Geographic Information Systems (GISs).

Spatial Data Warehouses (SDWs) combine DWs and SDBs for managing significant amounts of historical data that include spatial location. Merging these two technologies allows to exploit the capabilities of both systems for improving data analysis, visualization, and manipulation. DWs offer efficient access methods and management of high volumes of data. On the other hand, SDBs have a long experience in managing spatial data, and there is extensive research referring to spatial index structures, storage management, and dynamic query formulation. The experience gained in managing aggregated data in OLAP systems has already been extended for spatial data in spatial OLAP (SOLAP) systems. Further, since spatial data usually change over time, these changes can be represented using the time dimension provided by current DWs Malinowski (2004).

A multidimensional model is widely used in DWs and in OLAP systems for expressing users' requirements and to facilitate itôs afterward implementation. Extending a multidimensional model by the inclusion of spatial data provides a concise and organized SDW representation. It facilitates the delivery of data for SOLAP systems, spatial data mining, and spatial statistical analysis. Further, since it is platform independent, a conceptual multidimensional model allows establishing a communication bridge between users and designers. It reduces the difficulties of modeling spatial applications, since decision-making users do not usually possess the expertise required by software currently used for managing spatial data Malinowski (2004).

6 Visualisation techniques

The combination of data warehouse, data mining and data visualization is gradually becoming an indispensible organizational weapon for achieving competitive advantage in many data-driven industries Marakas (2003). Nabney et al (2005) indicated that data structural features can be effectively recognized by data seekers using data visualization. Data visualization is the process by which textual or numerical data are converted into meaningful images Marakas (2003). The reason why the data visualization can help on data mining is that the human brain is very effective in recognizing large amounts of graphical representations Ware (2004). Hence, if the visualization techniques can correctly convert the raw data into visual graphs, users can very likely detect the patterns hidden in text and numbers.

Yeh (2006) has also made comparisons of the visualization techniques, those are, Tree shaped diagrams and tree maps, parallel coordinates, Scatter-Plot Matrices, Survey plots as well as spatial/geographic visualization. For hierarchical datasets, treemap can catch the relationship immediately and present the linkage easily, so is the case for spatial/geographical datasets and spatial visualization technique. However, for general multidimensional dataset which is common in business data sources, the effectiveness of these two methods is falling behind the other three techniques. Among parallel coordinates, scatter-plot matrices, and survey plots, scatter-plot matrices is more recommendable. It performs well both in exploration and confirmation tasks, while maintaining its usefulness in presentation task and under pre-processing phase of data-mining task. Survey plot is ahead of parallel coordinates only when assisting pre-processing phase in a data mining job, but is equal to parallel coordinates at all other situations.

Miller (2004) stated visualization is a powerful strategy for leveraging the visual orientation of sighted human beings. Sighted humans are extraordinarily good at recognizing visual patterns, trends and anomalies; these skills are valuable at all stages of the knowledge discovery. Visualization can be used in conjunction with OLAP to aid the userôs synoptic sense of the database. Visualization can also be used to support data preprocessing, the selection of data mining tasks and techniques, interpretation and integration with existing knowledge Keim and Kriegel (1994). The approach that Stats SA took is an effort to enhance knowledge discovery is to build data marts and OLAP cubes using the survey data from QLFS.

7 Creation of data marts and the use of OLAP as a method on information delivery

Rob and Coronel (2000) describes data mart as a small, single-subject data warehouse subset that provides decision support to a small group of people. In CPM white paper they describe data mart consolidation as the process of centralising data from multiple, disparate sources into single, centralized enterprise data warehouse (EDW) that can be accessed by anyone in the organization who needs the information.

SAS describes OLAP cube as a set of data that is organized and structured in a hierarchical, multidimensional arrangement, often with many dimensions and levels of data. The way OLAP data is stored makes it readily available for detailed queries and analysis. The OLAP cube enables you or others to dynamically analyse data that has been summarized into multi-dimensional views and hierarchies. Rob and Coronel (2000) describe OLAP as decision support systems tools that use multi-dimensional data analysis techniques.

Cooper and Schindler (2003) see the advantage of multidimensional analysis, another phrase for OLAP being it provides fast, flexible data summarisation, analyses, and reporting capabilities with the ability to view trends over time. An example can be, how many households did not responded in a survey from City of Cape Town, compared to the same month last year.

By viewing aggregated data on multiple dimensions, both analysts and the end-user gain a deeper, more intuitive understanding of data in picture form. A multi-dimensional database typically contains three axes: (1) dimensions, like the field in a table; (2) measures, aggregate computations to be viewed; and (3) hierarchies, which impose structure on the dimensions for example time based hierarchy will be month, quarter, year Cooper and Schindler (2003).

8 The empirical study: demonstrating a component of knowledge discovery and its application in Statistics South Africa, which is Visualisation specifically using the OLAP cubes.

The Online Analytical Processing (OLAP) is currently in a process of implementation within the household survey methodology in Stats SA. The purpose of the solution is to empower the users to be able to do analysis from their location without assistance of other technical users. One of the objectives is to make results and information needed to monitor the survey readily available on a high level to different users within the organization.

One area where the OLAP system will come handy is in consolidating survey quality indicators which are not currently available in a single view. Among the indicators, we have; response rates, slippage rates, incoming and outgoing error rates in processing, imputation rates, master sample usage as well as master sample accountability and sampling review. We aim to provide the up-to-date information with regard to the trends of the above indicators.

8.1 Organisational problem

Statistics South Africa conducts household based and establishment based surveys on a regular basis. The surveys are designed using standard international methodologies. There are also several indicators of the surveys that are produced when data and survey results are released. Emphasis in the publication is on the statistical results. There is a multitude of information available in statistical datasets that is not analysied to inform continuous improvement process of studies as well as the decision making. In order to turn these data into knowledge they should be organised as a specific subject for example, response rates, slippage rates, unemployment rate etc.

The organisation of data under certain subject to answer a specific question is referred to as data mart. Furthermore we look at organising data in a multi-dimensional view to gain more understanding on the data. Surveys are conducted over a period of time in the organisation and they have time dimension. Data is collected and reported nationally, provincially, and for metro and non-metro district council that resembles location dimension.

The organisation uses SAS to analyse data and not all users are able to use SAS. The tool is aimed at both SAS users and non-SAS Users. Most users rely on the methodologists and survey statisticians for data provision on certain subjects which extends their time to get simple results. Due to simplicity of analysing data, users will be empowered to produce reports from these multi-dimensional pre-summarised data.

8.2 Proposed solution

The approach makes use of the visualisation as a tool in knowledge discovery. We demonstrate how visualisation techniques such as OLAP systems can be used in knowledge discovery. There are several ways in which the multi-dimensional data marts can be accessed. Users may use software that they have available for analysis and data manipulation in their standalone machines. There are other software available to organise data in a multi-dimensional way. One application is Online Analytical Processing tools. We consider using SAS OLAP cube studio to present data to variety of users within the organisation.

8.3 Creation of data marts

We are using Quarterly Labour Force Survey (QLFS) as an example of what kind of data marts are produced in the organisation. The system of creating data marts and OLAP cubes has a room to bring together different surveys in the organisation under a certain topic in one view. Among the data marts that are going to be produced in the organisation we have survey responses mart and response rate mart and Slippage rate mart.

1.1.1 Response Rate Mart

The response rate and its complement, the nonresponse rate, is commonly used as an indirect measure of quality of survey data; Executive Office of the President of the United States (2001). The computation of the response rates should be straight forward: the number of responding units divided by the total number of eligible units. These two options we consider to calculate non-response rates, those are, and firstly unweighted which is basically the number of interviews with reporting units divided by the number of eligible reporting units in the sample. It provides an indicator of the quality of the data collection and may be calculated at national, regional, and interviewer levels or for certain domains to evaluate performance.

Unweighted Response rate: $100 * \left(\dfrac{response}{response + nonresponse} \right)$

Secondly the weighted response rate which takes into account the different selection probabilities associated with different units in the sample. Weighted response rates indicate the proportion of the population (or some calculated subpopulation) that respond to the survey, and can be useful for an analystôs evaluation of the effect of the nonresponse on the survey estimates (Kasprzyk and Kalton 1997; Madow, Nisselson, and Olkin 1983). Some surveys use unweighted nonresponse to monitor the progress and analyse weighted nonresponse to determine whether some portions of the populations are underestimated.

Weighted Response Rates: $100 * \left(\frac{\sum w_i response}{\sum w_i \left(response + nonrespons\ e \right)} \right)$

Although there are other non-response rates options we are only considering the two options described above.

1.1.2 Slippage Rate Mart

Slippage rate for a population group is defined as the difference between the demography estimate that was used as benchmark for calibration weighting and the survey estimate obtained without the weight calibration. The slippage will also be referred to as under-coverage. There are two sources of slippage: (1) missed eligible households, and (2) missed eligible persons within the enumerated households.

Before analyzing the slippage rates, we should mention that the calibrated weights are constructed using only the population estimates as control totals. But, the household counts (e.g. number of households by household size, etc.) are not used as control totals. Therefore, the estimates of the number of households would be subject to sampling variability. Moreover, the estimates of number of households could be subject to potential bias because of slippage. The two sources of slippage would have different impact on the estimates of number of households. We discuss these two sources of slippage and their impact on the estimates of number of households and average household size.

The demography estimates of population as control totals are available for the cross-classification Age-Group (1) by Race by Gender at the national level, and by Age-Group (2) for each of the nine provinces. Slippage rate is the difference between the population estimate obtained from the survey (based on the design weights) and the known population used as control total expressed as percent of the known population total. Slippage rate can be estimated for any aggregation of the cells used as control totals for Calibration Weighting. Let $\hat{P}_{pg}$ be the population estimate for the province p and population sub-group g from the survey (based on the design weights), and $\tilde{P}_{pg}$ be the corresponding population control total from demography, then the slippage rate will be given by:

$$SR_{pg} = 100 \times \left(1 - \frac{\hat{P}_{pg}}{\tilde{P}_{pg}} \right)$$

The graphs of slippage rate are produced at the national level by plotting the slippage rate for the current quarter and the past quarters. This would provide an overall trend of the slippage rate.

8.4 Implementation and results presentation

We used data marts for the above two indicators those are response rates and slippage rate. The cubes are built in SAS OLAP cube studio that includes defining the dimensions and hierarchies as part of the cube structure. The results are viewed in Microsoft excel as well as the SAS Enterprise guide. We will demonstrate with an example what kind of intelligence can be derived from the trend produced using slippage rate cube. The hypothesis can be supposedly looking at the difference in magnitude of under-coverage between age groups. We test if there is difference between the slippage rates of age group 15 to 34 and the average slippage rates of other age groups. It must be known that the users may formulate various hypotheses and use other approaches to prove the hypothesis.

From our example we find that in the trend of slippage rates persons from 15 to 64 are consistently missed in surveys. We also find that of those that are missed, males are more likely to be missed than females. Studies have shown that the trend is evident world-wide that youth are among the hard-to-count population for various reasons that may differ from country to country.

9 Conclusions

The main objective of the study is to provide the quality indicators for quality improvement in surveys. Firstly we ensure that standard statistical methodologies are available to produce indicators. We have described the statistical approaches for providing indicators, which include response rate and slippage rate calculations. Secondly we considered presentation of indicators using visualisation tools in collaboration with the OLAP system. The third objective is to determine what kind of knowledge can be derived from visualisation, implemented in OLAP system. Lastly in addition to tested approaches and known indicators we consider traditional and geographic knowledge discovery approaches available to enhance our decision support systems.

We can turn statistical data into knowledge by simply organising it in a multi-dimensional structure and report using OLAP cubes. A cube can be viewed by multiple users without interfering with the source data. Third party option of using MS Excel to view cubes allows non SAS users to do independent analysis. The responsibility of educating users to analyse quality indicators lies in the hands of the survey methodologists. Having looked at the need for the information to support decision making the approaches addressed provide the most viable alternative to knowledge discovery those are, the decision support systems (DSS), knowledge discovery, Geographic knowledge discovery (GKD) and spatial data mining, Visualisation techniques, spatial representation of conceptual multidimensional models.

References

Alfred R (2005) *Knowledge Discovery: Enhancing Data Mining and Decision Support Intergration.* Artificial Intelligence Group, Department of Computer Science, The University of York, United Kingdom.

Choudry G.H (2007). *Sampling and weighting System.* Statistics South Africa South Africa.

Cooper D.R and Schindler (2003) *Business research methods.* Eighth edition. McGraw-Hill Irwin.

Executive Office of the President of the United States (2001). *Measuring and reporting Sources of Errors in Surveys. Statistical* Policy Working Paper Page 31.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). *From data mining to knowledge discovery: An overview,* in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining,* Cambridge, MA: MIT Press, 1-34.

Frawley et. al., 1991 W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. 1991 *Knowledge Discovery in databases: An Overview*. In G. Piatetsky-Shapiro and

Kamel S (1997). *DSS to Support Socio-Economic Development in Egypt. The American University in Cairo*.

Kasprzyk, D. and Kalton, G. 1997. *Measuring and reporting the quality of Survey Data.* Proceedings of Statistics Canada Symposium 97: New Directions in Surveys and Censuses. Ottawa: Statistics Canada. 179-184.

Keim, D. A. and Kriegel, H.-P. (1994). *Using visualization to support data mining of large existing databases,* in J. P. Lee and G. G. Grinstein (eds.) *Database Issues for Data Visualization,* Lecture Notes in Computer Science 871, 210-229.

Khorshid M (2004) *Model-Centered Government Decision Support Systems for socioeconomic development and in the Arab World.* Brussels, Belgium.

Nadeem M and Jaffri S.A.H (2000). *Application of Business Intelligence In the Banks (Pakistan).* Karachi, Pakistan.

Madow, W., Nisselson, H., and Olkin, I. 1983. Incomplete Data in Sample Surveys. New York: Academics Press.

Malinowski E and Zimanyi E (2004). *Representing Spatiality in a Conceptual Multidimensional Model.* Department of Computer & Network Engineering Universit ́e Libre de Bruxelles, Brussels, Belgium.

Marakas, G.M (2003). *Modern data warehousing, mining and visualization.* New Jersey: Prentice Hall.

Miller H.J. (2004). *Geographic Data Mining and Knowledge Discovery.* Department of Geography University of Utah. J. P. Wilson and A. S. Fotheringham (eds.) Handbook of Geographic Information Science.

Nabney, I.T., Sun, Y., Tino, P., & Kaban, A. (2005). Semi-supervised learning of hierarchical latent trait models for data visualization. IEEE Transactions on Knowledge and data Engineering, 17(3), 384-400

Rob P and Coronel C (2000). Database systems, *design implementation and management. Fourth edition.* Thomson Learning.

Shekhar, S. Zhang, P., Huang, Y. and Vatsavai, R. (2003) *"Trends in spatial data mining,"* in H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.), Data Mining: Next Generation: Challenges and Future Directions, AAAI/MIT Press, in press

Ware, C. (2004). *Information visualization: Perception for design* (2nd ed.). San Francisco: Morgan Kaufmann.

Yeh R.K. (2006) *Visualization Techniques for data mining in Business Context: A Comparative Analysis.* University of Texas at Arlington.